DOCUMENT RESUME

ED 368 771 TM 021 200

AUTHOR Thompson, Bruce
TITLE Common Methodology Mistakes in Dissertations, Revisited.
PUB DATE Apr 94
NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-6, 1994). For related document, see ED 301 595.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Behavioral Science Research; Case Studies; *Doctoral Dissertations; Error Patterns; Higher Education; *Information Dissemination; Literature Reviews; Measurement Techniques; Multivariate Analysis; Research Design; *Research Methodology; Research Problems; Research Reports; Statistical Significance; Test Interpretation; Test Reliability; Test Use; *Writing (Composition)
IDENTIFIERS Research Training; Stepwise Regression

ABSTRACT

Dissertations are an important component of the effort to generate knowledge. Thus, dissertation quality may be seen by accreditation and coordinating-board reviewers as a noteworthy reflection on the quality of doctoral programs themselves. The present study reviews methodological errors within Ph.D. dissertations. The illustrative errors are presented within the framework of seven analytic principles, each of which is explored in some detail. These are (1) Because tests are not reliable, but data are, data must generally be evaluated for measurement integrity; (2) Statistical significance testing is of limited importance and should be augmented by other analyses; (3) Multivariate methods are usually vital in behavioral research; (4) Result interpretations should not be based only on standardized weights; (5) Intervally scaled variables should generally not be converted to the nominal level of scale; (6) Covariance corrections are generally either unnecessary or ineffective and should usually be avoided; and (7) Stepwise methods should not be used. Includes six tables, one figure. (Contains 83 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

dissrnew.wp1 3/9/94

ED 368 771

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BRUCE THOMPSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Common Methodology Mistakes in Dissertations, Revisited

Bruce Thompson

Texas A&M University 77843-4225
and
Baylor College of Medicine

Paper presented at the annual meeting (session #24.58) of the American Educational Research Association, New Orleans, April 6, 1994.

TM021200

BEST COPY AVAILABLE

ABSTRACT

Dissertations are an important component of the effort to generate knowledge. Thus, dissertation quality may be seen by accreditation and coordinating-board reviewers as a noteworthy reflection on the quality of doctoral programs themselves. The present study reviews methodological errors within Ph.D. dissertations. The illustrative errors are presented within the framework of seven analytic principles, each of which is explored in some detail.

> Ruggiero & [sic] Enyart (1987) examined the construct and discriminant validity of the CDI...." (p. 17)

Nor does the practice of making "assumptions" regarding critical design features fall within the present discussion. Some students seem to feel that one can "assume away" a burden of proof establishing that a study's design and measurement features were sufficiently sound to inform meaningful result interpretation. ##Thurstone (1991) provides an example in a section of that study titled "Assumptions Regarding the Study", and consisting of exactly two sentences:

> 1. The two hundred and forty children selected as participants in this study are representative of Hispanic children in South Texas. 2. Testing was conducted in the same manner with all the children and their parents or guardians." (p. 7)

One wonders what import the study's results would have if either of these assumptions writ as wishes were, in actuality, unfulfilled. It is appropriate to offer assumptions or postulates in studies, but only when one can present some theoretical or empirical evidence that these assumptions are at least likely to be true.

Finally, the rampant tendency of these students in their Discussion chapters to not cite the <u>specific</u> statistics (or at least the specific tables containing the statistics), upon which their interpretations are based, will not be discussed in detail here. ##Stephenson (1992) provides a compelling example of this ilk.

##Stephenson (1992) reported in Chapter IV a series of univariate one-way ANOVA <u>F</u> tests, a discriminant function analysis, and a host of bivariate correlation coefficients. In Chapter V ##Stephenson did not directly address any of the statistics evaluated by these various analyses, nor were there specific references in the Chapter V to any of the tabled statistics in Chapter IV. Thus, the reader is forced to guess the basis of these Chapter V conclusions, or to presume that the analyses and the conclusions were simply unrelated.

<u>Principle #1</u>: *Because tests are <u>not</u> reliable, scores are, the data in hand must generally be evaluated for measurement integrity.*

<u>Tests are Not Reliable or Unreliable</u>. Too few researchers act on a conscious recognition that *reliability is a characteristic of <u>scores</u> or the data in hand*. Test booklets are not impregnated with reliability during the printing process. The same WISC-R that yields reliable scores for some adults on a given occasion of measurement will not necessarily do so when the same test is administered to first-graders.

Many researchers recognize these dynamics on some level, but paradigm influences constrain some researchers from actively integrating this presumption into their actual analytic practice. The pernicious practice of saying, "the test is reliable", creates

a language that predisposes researchers against acting on a conscious realization that tests themselves are not reliable, and acting accordingly (Thompson, 1994).

As Rowley (1976, p. 53, emphasis added) argued, "It needs to be established that an instrument itself is neither reliable nor unreliable.... A single instrument can produce *scores* which are reliable, and other *scores* which are unreliable." Similarly, Crocker and Algina (1986, p. 144, emphasis added) argued that, "...A test is not 'reliable' or 'unreliable.' Rather, reliability is a property of the *scores* on a test for a *particular* group of examinees."

In another widely respected text, Gronlund and Linn (1990, p. 78, emphasis in original) noted,

> Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument itself.... Thus, it is more appropriate to speak of the reliability of the "test scores" or of the "measurement" than of the "test" or the "instrument."

And Eason (1991, p. 84, emphasis added) argued that:

> Though some practitioners of the classical measurement paradigm [incorrectly] speak of reliability as a characteristic of tests, in fact reliability is a characteristic of *data*, albeit data generated on a given measure administered with a given protocol to given subjects on given occasions.

The subjects themselves impact the reliability of scores, and thus it becomes an oxymoron to speak of "the reliability of the test" without considering to whom the test was administered, or other facets of the measurement protocol. Reliability is driven by variance--typically, greater score variance leads to greater score reliability, and so more *heterogeneous* samples often lead to more *variable* scores, and thus to higher reliability. Therefore, the same measure, when administered to more heterogenous or to more homogeneous sets of subjects, will yield scores with differing reliability. As Dawis (1987, p. 486) observed, "...Because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population--an obvious but sometimes overlooked point."

Our shorthand ways of speaking (e.g., language saying "the test is reliable") can itself cause confusion and lead to bad practice. As Pedhazur and Schmelkin (1991, p. 82, emphasis in original) observed, "Statements about *the* reliability of a measure are... inappropriate and potentially misleading." These telegraphic ways of speaking are not inherently problematic, but they often later become so when we come unconsciously to ascribe literal truth to our shorthand, rather than recognizing that our jargon is sometimes telegraphic and is not literally true. As noted elsewhere:

> This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy

> speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice. Thompson (1992c, p. 436)

One sloppy practice is not calculating, reporting, and interpreting the reliability of one's own scores for one's own data. As Pedhazur and Schmelkin (1991, p. 86, emphasis in original) argued:

> Researchers who bother at all to report reliability estimates for the instruments they use (many do not) frequently report only reliability estimates contained in the manuals of the instruments or estimates reported by other researchers. Such information may be useful for comparative purposes, but it is imperative to recognize that the *relevant reliability estimate is the one obtained for the sample used in the [present] study under consideration*.

Why Score Reliability in So Important. In one book exploring the intimate linkages between measurement error variance and our attributions about the origins of variance in our substantive basic or applied research, Pedhazur and Schmelkin (1991) noted,

> Measurement error is the Achilles' heel of sociobehavioral research. Although most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned. Thus, many students get the impression that no special competencies are necessary for the development and use of measures... (pp. 2-3)

Therefore, it should not be surprising that studies of research reports in journals indicate insufficient attention is paid to the impacts of measurement integrity on the integrity of substantive research conclusions. For example, with respect to the American Educational Research Journal, Willson (1980) reported that:

> ...Only 37% of the AERJ studies explicitly reported reliability coefficients for the data analyzed. Another 18% reported only indirectly through reference to earlier research.... That reliability... is unreported in almost half the published research is... inexcusable at this late date.... (pp. 8-9)

A more recent "perusal of contemporary psychology journals demonstrates that quantitative reports of scale reliability and validity estimates are often missing or incomplete" (Meier & Davis, 1990, p. 113); and that "the majority [95%, 85% and 60%] of the scales described in the [three Journal of Counseling Psychology] JCP volumes [1967, 1977 and 1987] were not accompanied by reports of psychometric properties" (p. 115).

This state of affairs is surprising, given two related trends within the literature. First, since the influential articles by

Cohen (1968) and Knapp (1978) appeared, more researchers have recognized that all parametric statistical analyses are correlational (Thompson, 1991a), and that substantive variance-accounted-for effect sizes expressed as r^2 analogs can be interpreted in all studies. Second, the importance of interpreting effect sizes as against statistical significance tests has been increasingly recognized (e.g., Thompson, 1993b), as reflected, for example, in a recent procession of articles within the American Psychologist (cf. Cohen, 1990; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989).

Nevertheless, too few researchers act on the premise that score reliability establishes a ceiling for substantive effect sizes. These impacts can be readily illustrated in a concrete example using the bivariate correlation as an heuristic.

It has been recognized in textbooks dating back to the 1950s, and in more recent books as well (e.g., Pedhazur & Schmelkin, 1991, p. 114), that a correlation coefficient "corrected" for attenuation due to measurement error ($\hat{r}_{XY}$) can be estimated as:

$$\hat{r}_{XY} = r_{XY} / [(r_{XX} * r_{YY})^{.5}],$$

where r_{XY} is the calculated bivariate relationship between scores on variables X and Y, and r_{XX} and r_{YY} are respectively the reliability coefficients for scores on X and Y. This algorithm can be re-expressed in the more familiar metric of common variance, as is often done in popular variance-accounted-for effect size statistics (e.g., r^2, R^2, eta^2, omega^2):

$$\hat{r}_{XY}^2 = r_{XY}^2 / (r_{XX} * r_{YY})$$

Through algebraic manipulation, the detectable effect size, given knowledge of "true" relationship, $\hat{r}_{XY}^2$, and the reliabilities of the two sets of scores, is:

$$r_{XY}^2 = \hat{r}_{XY}^2 * (r_{XX} * r_{YY})$$

Even if the "true" relationship between perfectly reliable measures of X and Y was perfect, i.e., $\hat{r}_{XY}^2 = 1.0$, the detectable effect in any study can *never exceed the product of the reliability coefficients for the two sets of scores*:

$$r_{XY}^2 = 1 * (r_{XX} * r_{YY})$$

For example, even when $r_{XY}^2 = 1.0$, if both sets of scores have reliability coefficients of .7, the detectable effect cannot exceed .49. Clearly, measurement error prospectively impacts the effect size that we can obtain in a planned study and also should be retrospectively considered when interpreting calculated effects once the study has been done.

The failure to consider score reliability in substantive research may exact a toll on the interpretations within research studies. We may conduct studies that could not possibly yield noteworthy effect sizes. Or we may not accurately interpret our results if we do not consider the reliability of the scores we are actually analyzing.

These practices may be caused by misperceptions that tests can be reliable or valid. These misperceptions themselves may be caused, or are at least reinforced, by the use of telegraphic language that comes to be unconsciously believed as literal truth,

and then unconsciously incorporated into paradigms for behavior.

Examples of Bad Practice Within Dissertations. Some students are to be commended for computing reliability coefficients for the scores for their own data. For example, ##Eysenck (1992) reported that:

> A Chronbach [sic] alpha reliability coefficient for this measure was computed from the data for the current study and found to be .71. (p. 60)

However, it is very unfortunate that almost all students used *bad language* suggesting that they believed reliability inures as a characteristic to tests themselves. For example, ##Fisher (1992) noted,

> The scores of the entire research sample (N = 234) was [sic] used in determining the reliability of the scale [sic] which resulted in coefficient alpha [sic] of .84. (pp. 106-107)

Still more troubling is the behavior of calculating the reliability of one's own scores and then still *conducting substantive analyses with scores of acknowledged questionable reliability* (and thus questionable validity as well). ##Velicer (1992) provides an example, prior to the substantive analyses including scores on three scales:

> The researcher of this study also conducted a Cronbach alpha reliability analysis on the [three] subscales. The reliability coefficients for each subscale and total questionnaire [sic] are as follows: Task Orientation, r = .93; Adaptability, r = .66; and Reactivity r = .45. (pp. 44-45, emphasis added)

Even more troubling is the tendency to *not fully evaluate the reliability of the scores that are actually being evaluated in substantive analyses*. This may occur even when the researchers themselves suggest that there was, in fact, every reason to believe that their own scores were not reliable. For example, ##McDonald (1992) reported:

> Catecholamines are notoriously variable both between individuals and within the same individual under varying conditions. Their production and excretion can be affected by many physical, physiological and psychological factors. Variables that may have influenced catecholamine levels during the 24 hour collection include: diet, temperature and season of the year, mental activity or work, physical activity or work, and emotional state (anxiety, aggression, fear). (pp. 59-60)

It is especially intriguing when students (a) argue that reliability analyses are needed, (b) have the data to conduct such analyses in their own studies, but (c) do not conduct the analyses that they themselves argue are needed. For example, ##Cattell (1992) noted that:

> Although reliability and validity data are unavailable for the GDS/Amended, the differences in

> wording of the four changed items appear slight. It is thought [i.e., assumed] that studies on the original GDS give [sic] adequate reliability and validity for the GDS/Amended. (p. 46)

Subsequently, ##Cattell (1992, p. 91) offered a recommendation for future studies, correctly arguing that, "Reliability and validity studies are needed on the amended assessments for nursing homes and/or institutions..., as none were found in a review of the literature."

Most troubling of all is a pattern where students do <u>not</u> analyze the reliability of their scores, *when their scores are actually not very reliable, and these problems are not considered during analysis and/or interpretation*. For example, though ##Huberty (1991, p. 63) did not report an alpha coefficient, sufficient information was provided that this estimate could be computed by the reader. The relevant calculations are:

$$\alpha_X = [v / (v-1)] [1 - ((\Sigma SD_i^2) / SD_X^2)]$$
$$[4 / (4-1)] [1 - ((.32^2 + .37^2 + .50^2 + .33^2)/1.04^2)]$$
$$[1.333333] [1 - ((.1024 + .1369 + .2500 + .1089)/1.0816)]$$
$$[1.333333] [1 - ((.1024 + .1369 + .3589) / 1.0816)]$$
$$[1.333333] [1 - ((.1024 + .4958) / 1.0816)]$$
$$[1.333333] [1 - (.5982) / 1.0816)]$$
$$[1.333333] [1 - .553069]$$
$$[1.333333] [.446930]$$
$$+.595907$$

But ##Lawley (1992, p. 59) provides the most stunning example of all! Though not computed, sufficient information was available in the report to calculate a rough KR_{21} estimate of reliability for one set of scores on PNID, an important variable in that study:

$$KR_{21} = [v / (v -1)] [1 - (\bar{X} (v - \bar{X}))/(v (SD^2))]$$
$$[13/(13-1)] [1 - (1.18 (13 - 1.18)) / (13 (0.7^2))]$$
$$[13 / 12] [1 - (1.18 (11.82)) / (13 (0.49))]$$
$$[1.083333] [1 - (13.9476) / (6.37)]$$
$$[1.083333] [1 - (2.189576)]$$
$$[1.083333 [-1.18957]$$
$$\mathbf{-1.28870}$$

Reinhardt (1991) provides an excellent review of reliability coefficients, and the factors that impact score reliability. Certainly, it would be difficult to be sanguine about ##Lawley's (1992) conclusions, given an estimated reliability coefficient of this magnitude.

<u>Principle #2</u>: *Statistical significance testing is of limited importance, especially as regards measurement characteristics statistics, and should be augmented by other analyses.*

<u>What Statistical Significance Testing Is and Isn't</u>. Science is about the business of identifying relationships that recur under stated conditions. Unfortunately, too many researchers, consciously or unconsciously, incorrectly assume that the <u>p</u> values calculated in statistical significance tests evaluate the

probability that results will replicate (Carver, 1978, 1993). Such researchers often explain what p calculated is by invoking vague and embarrassing anorphisms such as, "p calculated (or statistical significance testing) evaluates whether results were 'due to chance'".

It is true that statistical significance tests do focus on the null hypothesis. It is also true that such tests evaluate sample statistics (e.g., *sample* means, standard deviations, correlations coefficients) in relation to unknowable population parameters (e.g., *population* means, standard deviations, correlations coefficients).

But far too many researchers incorrectly interpret statistical significance tests as evaluating *the probability that the null is true in the population, given the sample statistics for the data in hand*. This would, in fact, be a very interesting issue to evaluate.

If p calculated informed the researcher about the truth of the null in the population, then this information would directly test the replicability of results. Assuming the population itself remained stable, future samples from the population, if representative, should yield similar results. In this case, results for which the null was found to not be true in the population would therefore be likely to be replicated in future samples from the same population where the null would also likely be rejected. Unfortunately, this is not what statistical significance tests, and not what the associated p calculated values evaluate.

It is true that the p(robability) values calculated in statistical significance testing, which range from 0 to 1 (or 0% to 100%), do require that a "given" regarding the population parameters must be postulated. The characteristics of the population(s) directly affect what the calculated p values will be, and are considered as part of the calculations of p.

For example, if we draw two random samples from two populations, both with equal means, then the single most likely sample statistics (i.e., the sample statistics with the *largest p calculated value*) will be two equal sample means. These sample results are the most likely for these populations. But these exact same sample statistics would be less likely (i.e., would yield a *smaller p calculated value*) if the two populations had parameter means that differed by one unit. And the sample statistics involving exactly equal sample means would be still less likely (i.e., would yield a still *smaller p calculated value*) if the two population means differed by two units.

Indeed, specific population parameters must unavoidably be assumed even to determine what the p calculated is for the sample statistics. Given that population parameters directly affect the calculated p(robability) of the sample statistics, one must assume particular population parameters associated with the null hypothesis being tested (e.g., specific means, medians, standard deviations, correlation coefficients), because there are infinitely many possibilities of what these parameters may be in the

population(s).

Only by assuming specific population parameters can a single answer be given to the question, "what is the p(robability) of the sample statistics, assuming the population has certain parameters?" Without the assumption of specific population parameters being true in the population, there are infinitely many plausible estimates of p, and the answers to the question actually posed by statistical significance testing become mathematically indeterminate.

Classically, to get a single estimate of the p(robability) of the sample statistics, the null hypothesis is posited to be exactly true in the population. Thus, statistical significance testing evaluates *the probability of the sample statistics for the data in hand, given that null hypothesis is presumed to be exactly true as regards the related parameters in the population.*

Of course, this p is a very different animal than one which evaluates the probability of the population parameters themselves, and the statistical significance testing logic itself means that p evaluates something considerably less interesting than result replicability. As Shaver (1993) recently argued so emphatically:

> [A] test of statistical significance is not an indication of the probability that a result would be obtained upon replication of the study. A test of statistical significance yields the probability of a result occurring under [an assumption of the truth of] the null hypothesis [in the population], not the probability that the result will occur again if the study is replicated. Carver's (1978) treatment should have dealt a death blow to this fallacy.... (p. 304)

Furthermore, the requirement that statistical significance testing presume an assumption that the null hypothesis is true in the population is a requirement that an untruth be posited. As Meehl (1978, p. 822) notes, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false." Similarly, Hays (1981, p. 293) points out that "[t]here is surely nothing on earth that is completely independent of anything else [in the population]. The strength of association may approach zero, but it should seldom or never be exactly zero."

One logic explaining why the null cannot be true in the population is mathematical. There are infinitely many possible parameters (e.g., means, standard deviations) in the population(s). Probability is the frequency of occurrence of an event divided by the total number of possible events. Therefore, the "point probability" of any single event (e.g., two populations with exactly equal means, a population with the parameter correlation coefficient exactly equal to zero) in the population is infinitely small. Thus, the probability of the null hypothesis being exactly or literally true in the population is infinitely small.

There is a very important implication of the realization that the null is not literally true in the population. The most likely sample statistics for samples drawn from populations in which the

null is not literally true are sample statistics which do not correspond to the null hypothesis, e.g., there are some differences in sample means, or r in the sample is not exactly 0. And whenever the null is not exactly true in the sample(s), then the null hypothesis will always be rejected at some sample size. As Hays (1981, p. 293) emphasizes, "virtually any study can be made to show significant results if one uses enough subjects."

Although statistical significance is a function of at least seven interrelated features of a study (Schneider & Darcy, 1984), sample size is a basic influence on significance. Thus, some researchers (Thompson, 1989a, 1993b) have advocated interpreting statistical significance tests only within the context of sample size. In any case, all this means that:

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. This tautology has created considerable damage as regards the cumulation of knowledge... (Thompson, 1992c, p. 436)

Thus, statistical significance testing can be a circuitous logic requiring us to invest energy to determine that which we already know, i.e., our sample size. And this energy is not invested in judging the noteworthiness of our effect sizes or the replicability of our effect sizes, since statistical significance testing does not evaluate these considerably more important issues. The recent Summer, 1993, special issue (Vol. 61, No. 4) of the Journal of Experimental Education provides a lucid and thorough treatment of these and related matters. Decades of effort "to exorcise the null hypothesis" (Cronbach, 1975, p. 124) continue.

Examples of Bad Practice Within Dissertations. Five areas of bad practice as regards statistical significance testing were isolated within these dissertations. First, *essentially all students in their dissertations always said "significant" when they meant "statistically significant"*. Clearly, statistical significance does not evaluate result importance, and always using the phrase "statistically significant" when referring to inferential tests helps at least a little to avoid confusing statistical significance with result importance. As Thompson (1993b) emphasizes:

> Statistics can be employed to evaluate the probability of an event. But importance is a question of human values, and math cannot be employed as an atavistic escape (a la Fromme's Escape from Freedom) from the existential human responsibility for making value judgments. If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating p's, and so p's cannot be blithely used to infer the value of research

> results. Like it or not, empirical science in inescapably a subjective business. (p. 365)

Thus, some researchers strongly admonish against the use of only the words "significant" or "significance", when referring to statistical significance (e.g., Carver, 1993, p. 288). Similarly, Moore (1991), a doctoral student, argued in an AERA paper she presented:

> First, many times the word "statistically" is not used in the description of the [statistically significant] results. Researchers may report that "a significant difference was obtained." This statement does imply the word "statistically," but many people may not understand the implication or hint regarding what is being discussed. (p. 1)

In this vein, ##Mulaik (1992) presents a cascade of examples of bad language practice, each and all ignoring Moore's (1991) admonitions (perhaps these two doctoral students never met during their concurrent enrollment):

> The ANOVA yielded a significant [sic] result... (p. 59)
>
> This significant [sic] result can be seen in Table 3... (p. 59)
>
> The ANOVA yielded a significant [sic] result... (p. 59)
>
> The significant [sic] result can be seen... (p. 61)
>
> ...rated the counselor significantly [sic] higher... (p. 61)
>
> The ANOVA failed to yield a significant [sic] result.... (p. 61)
>
> The ANOVA indicated that the interaction... was significant [sic]... (p. 61)
>
> The ANOVA failed to yield significant [sic] results... (p. 63)
>
> The interaction... also failed to yield a significant [sic] effect... (p. 63)
>
> The results indicated a significant [sic] positive relationship... (p. 68)

Second, *too few dissertation researchers reported and interpreted effect sizes.* With respect to effect sizes, very few of the students calculated either uncorrected (e.g., $\underline{r}^2$, $\underline{R}^2$, eta^2) or corrected (e.g., adjusted $\underline{R}^2$, omega^2) effect sizes (see Snyder & Lawson, 1993, for a very useful discussion). One commendable exception was ##Stephenson (1992), who employed the Wherry correction to "adjust" $\underline{R}^2$:

> Due to the high number of predictors (21) and the low number of cases (48), an estimation of the shrinkage of R^2 =.45 was calculated.... (p. 54)

Regrettably, when the correction to the calculated $\underline{R}^2$ of 45% was applied, to take into account the large number of variables relative to the small number of subjects, the variance-accounted-for effect size "shrunk" by 44.55 to .45%, i.e., became virtually zero. Then ##Stephenson (1992) offers two conclusions that are, respectively, (a) true and (b) difficult to understand:

> Therefore, the generalization of the predictor variables [sic] in the population is virtually impossible. However, it may be possible to generalize prediction in a referred population. (p. 54)

Third, *the dissertation researchers did not empirically evaluate the likely replicability of their results*. Since science is about the business of identifying results that replicate under stated conditions, and since statistical significance tests do <u>not</u> evaluate result replicability, it is especially important to evaluate the likely replicability of one's results. Identifying effects such as cold fusion is fun, and makes one popular at conventions, only until no one else can replicate your discovery.

Replicability can be evaluated best by actually replicating results. Alternatively, so-called "internal" replicability analyses can be conducted, using the data in hand in a single study, and applying analyses such as (a) cross-validation, (b) the jackknife, and (c) the bootstrap. Thompson (in press) and Reinhardt (1992) provide readable treatments of these strategies.

Fourth, *too few dissertation researchers consider the effects of sample size on statistical significance tests, either analytically or in subjective interpretations of results*. As noted previously, sample size is a basic influence on statistical significance tests. These effects can be directly evaluated using the so-called "what if" analyses recommended by Thompson (1989a, 1993b) and Snyder and Lawson (1993). Even when sample size is not considered in an empirical evaluation of statistical tests, at least these influences should be considered in subjective interpretations of results.

No dissertation students in these cohorts reported "what if" analyses for their tests of statistical significance. However, at least ##Mosteller (1991) acknowledged that these influences may account for differences in statistical significance tests of the same hypotheses across studies with different sample sizes:

> The larger sample size used by Ward and Meyers (1984) may have contributed to more statistically significant differences between correlations that were found with Rusk et al. (1979). Ward and Meyers (1984) used a sample size of 100, while Rusk et al. (1979) used a sample size of 32. (p. 70)

However, even ##Mosteller's (1991) subjective interpretation on this issue would have been improved by conducting empirical "what if" analyses to determine specifically whether differences in sample sizes could plausibly have accounted for the contradictory results in these two studies, given their actual specific effect

sizes and their actual n's.

Fifth, *dissertation students frequently conduct inappropriate statistical significance tests of reliability and validity coefficients*. Huck and Cormier (in press) explain why testing the statistical significance of reliability coefficients is not sensible:

> When statistically testing reliability coefficients, however, we question whether much is gained simply by saying that a test-retest correlation (or any other kind of reliability coefficient) is significantly different from zero. We say this because it is possible for a researcher to have a very low reliability coefficient turn out to be [statistically] significant, so long as the sample size is large enough.

Yet, just such tests are common within the dissertations studied here.

For example, it was commendable that ##Spearman computed reliability coefficients for the scores in that study, and especially commendable that analysis of an unreliable set of scores was "abandoned":

> ...[W]hen it was found that the currently-used ratio-of-change method of scoring the CAT (Probe Phase score/Acquisition Phase score) was unstable (test-retest reliability [for data in the present study] $r = .38$, $p < .004$), this score was abandoned. (p. 117)

But the statistical significance test of this reliability coefficient was unnecessary and inappropriate in this study. Note that the reliability coefficient was statistically significant at conventional alpha levels, even though ##Spearman (correctly) decided that the coefficient was unacceptably low.

##Cattell (1992) did not apparently make an equally incisive judgment to abandon an unreliable set of scores, perhaps because the "test-retest correlation" coefficient in that study was statistically significant, even though the two sets of scores shared only 28.09% ($r^2 = .53^2 = .2809$) of their variance in common:

> [The t]est-retest [reliability] correlation [involving a five week delay] for subjective Social Resource items on the Community Survey Questionnaire was $r = .53$, $p < .001$. (p. 51)

Especially intriguing is the study by ##Bartlett (1992), who reported that:

> This [null] hypothesis stated that the correlation between the GDS and the GDS-SF [Short Form] was zero. A Pearson correlation [concurrent validity] coefficient between the two scales was .94, $p<.001$. (p. 76)

A one-tailed statistical significance test of an r of roughly .94, even at the $\alpha=.01$ level of statistical significance, will be statistically significant with an n as small as 5! Statistical

tests of such coefficients in a measurement context clearly make little sense.

Principle #3: *Multivariate methods are usually vital in behavioral research.*

Two Reasons Why Multivariate Methods are Usually Vital. There are two reasons why multivariate methods are so important in behavioral research. These are elaborated by Fish (1988), and elsewhere, and are summarized here.

First, *multivariate methods limit the inflation of Type I "experimentwise" error rates.* It is clear that, "Whenever multiple statistical tests are carried out in inferential data analysis, there is a potential problem of 'probability pyramiding'" (Huberty & Morris, 1989, p. 306). And as Morrow and Frankiewicz (1979) emphasize, it is also clear that in some cases the inflation of experimentwise error rates can be quite serious.

Most researchers are familiar with "testwise" alpha. But while "testwise" alpha refers to the probability of making a Type I error for a given hypothesis test, "experimentwise" error rate refers to the probability of having made a Type I error anywhere within the study. When only one hypothesis is tested for a given group of people in a study, "experimentwise" error rate will exactly equal the "testwise" error rate. But when more than one hypothesis is tested in a given study with only one sample, the two error rates may not be equal.

Given the presence of multiple hypothesis tests (e.g., two or more dependent variables) in a single study with a single sample, the testwise and the experimentwise error rates will still be equal only if the hypotheses (or the dependent variables) are perfectly correlated. Logically, the correlation of the dependent variables will impact the experimentwise error rate, because, for example, when one has perfectly correlated hypotheses, in actuality, one is still only testing a single hypothesis. Thus, two factors impact the inflation of experimentwise Type I error: (a) the number of hypotheses tested using a single sample of data, and (b) the degree of correlation among the dependent variables or the hypotheses being tested.

When the dependent variables or hypotheses tested using a single sample of data are perfectly uncorrelated, the experimentwise error rate (α_{EW}) can be calculated. This is done using what is called the Bonferroni inequality (Love, 1988):

$$\alpha_{EW} = 1 - (1 - \alpha_{TW})^{K},$$

where k is the number of perfectly uncorrelated hypotheses being tested at a given testwise alpha level (α_{TW}).

For example, if three perfectly uncorrelated hypotheses (or dependent variables) are tested using data from a single sample, each at the $\alpha_{TW}=.05$ level of statistical significance, the experimentwise Type I error rate will be:

$$\begin{aligned} \alpha_{EW} &= 1 - (1 - \alpha_{TW})^{K} \\ &= 1 - (1 - .05)^{3} \\ &= 1 - (.95)^{3} \end{aligned}$$

$$
\begin{aligned}
&= 1 - (.95(.95)(.95)) \\
&= 1 - (\ .9025\ (.95)) \\
&= 1 - \quad .857375 \\
\alpha_{EW} &= .142625
\end{aligned}
$$

Thus, for a study testing three perfectly uncorrelated dependent variables, each at the α_{TW}=.05 level of statistical significance, the probability is .142625 (or 14.2625%) that one or more null hypotheses will be incorrectly rejected within the study. Most unfortunately, knowing this will not inform the researcher as to which one or more of the statistically significant hypotheses is, in fact, a Type I error. Table 1 presents these calculations for several conventional α_{TW} levels and for various numbers of perfectly uncorrelated dependent variables or hypotheses.

INSERT TABLE 1 ABOUT HERE.

But these concepts are too abstract to be readily grasped. Happily, Witte (1985, p. 236) explains the two error rates using an intuitively appealing example involving a coin toss. If the toss of heads is equated with a Type I error, and if a coin is tossed only once, then the probability of a head on the one toss (α_{TW}), and of at least one head within the set (α_{EW}) of one toss, will both equal 50%.

If the coin is tossed three times, rather than only once, the "testwise" probability of a head on each toss is still exactly 50%, i.e., α_{TW}=.50 (not .05). The Bonferroni inequality is a literal fit to this example situation (i.e., is a literal analogy rather than a figurative analogy), because the coin's behavior on each flip is literally uncorrelated with the coin's behavior on previous flips. That is, a coin is not aware of its behavior on previous flips and does not alter its behavior on any single flip given some awareness of its previous behavior.

Thus, the "experimentwise" probability (α_{EW}) that there will be at least one head in the whole set of three flips will be exactly:

$$
\begin{aligned}
\alpha_{EW} &= 1 - (1 - \alpha_{TW})^{K} \\
&= 1 - (1 - .50)^{3} \\
&= 1 - (\ .50\ \ )^{3} \\
&= 1 - (.50(.50)(.50)) \\
&= 1 - (\ .2500\ (.50)) \\
&= 1 - \quad .125000 \\
\alpha_{EW} &= .875000
\end{aligned}
$$

Table 2 illustrates these concepts in a more concrete fashion. There are eight equally likely outcomes for sets of three coin flips. These are listed in the table. Seven of the eight equally likely sets of three flips involves one or more Type I error, defined in this example as a heads. And 7/8 equals .875000, as expected, according to the Bonferroni inequality.

INSERT TABLE 2 ABOUT HERE.

Researchers control "testwise" error rates by picking small values, usually 0.05, for the "testwise" alpha. "Experimentwise" error rates can be limited by employing multivariate statistics to test omnibus hypotheses as against lots of discrete univariate hypotheses.

Paradoxically, although the use of several univariate tests in a single study can lead to too many null hypotheses being spuriously rejected, as reflected in inflation of the "experimentwise" error rate, it is also possible that the failure to employ multivariate methods can lead to a failure to identify statistically significant results which actually exist. Fish (1988) and Maxwell (1992) both provide data sets illustrating this equally disturbing possibility. This means that the so-called "Bonferroni correction" is <u>not</u> a satisfactory solution to this problem.

The "Bonferroni correction" involves using a new testwise alpha level, α_{TW}*, computed, for example, by dividing α_{TW} by the number of <u>k</u> hypotheses in the study. This approach attempts to control the experimentwise Type I error rate by reducing the testwise error rate level. However, the use of the "Bonferroni correction" does <u>not</u> address the second (and more important) reason why multivariate methods are so often vital, and so even with this correction univariate methods usually still remain unsatisfactory.

Multivariate methods are also often vital in behavioral research because, second, *multivariate methods best honor the reality to which the researcher is purportedly trying to generalize*. As noted previously, since statistical significance testing and error rates may not be the most important aspect of research practice (Thompson, 1989a, 1993b), this second reason for employing multivariate statistics is actually the more important of the two grounds for using these methods.

Implicit within all analyses is an analytic model. Each researcher also has a presumptive model of what reality is believed to be like. It is critical that our analytic models and our models of reality match, otherwise our conclusions will be invalid. It is generally best to consciously reflect on the fit of these two models whenever we do research. Of course, researchers with different models of reality may make different analytic choices, but this is not disturbing since analytic choices are philosophically driven anyway (Cliff, 1987, p. 349).

But Thompson (1986, p. 9) notes that the reality about which most researchers wish to generalize is usually one "in which the researcher cares about multiple outcomes, in which most outcomes have multiple causes, and in which most causes have multiple effects." Given such a model of reality, it is critical that the full network of all possible relationships be considered *simultaneously* within the analysis.

Conceptually, dependent variables can interact in a multivariate analysis, just as independent variables can interact in a multi-way ANOVA. This means that one can obtain statistically significant and large effect sizes in multivariate analyses of the same data yielding no statistically significant or large effect

sizes with univariate analyses, as Fish's (1988) example data illustrate. Thus, Tatsuoka's (1973, p. 273) previous remarks remain telling:

> The often-heard argument, "I'm more interested in seeing how each variable, in its own right, affects the outcome" overlooks the fact that any variable taken in isolation may affect the criterion differently from the way it will act in the company of other variables. It also overlooks the fact that multivariate analysis--precisely by considering all the variables simultaneously--can throw light on how each one contributes to the relation.

An Error: Using Univariate Tests in a Multivariate Context. In classical ANOVA, post hoc comparisons are necessary to determine which groups differ if (a) a statistically significant omnibus test is isolated and (b) there are more than two groups involved in the effect. But in multivariate analyses, such as classical MANOVA, when there is a statistically significant omnibus effect post hoc tests will be necessary to address either or both of two questions: (1) which groups differ?, and (2) on which dependent variables do groups differ?. Thus, even when there are only two groups in a multivariate analysis, a statistically significant omnibus result will still require post hoc exploration to address the second question, (2) on which dependent variables do groups differ?.

Too often researchers use MANOVA to test the full network of variable relationships, and if they obtain statistically significant results, then employ univariate ANOVAs or t-tests to do the post hoc work. This is the so-called "protected F-test" analytic approach.

The "protected F-test" analytic approach is inappropriate and wrong-headed. The multivariate analysis evaluates multivariate synthetic variables, while the univariate analysis only considers univariate latent variables. Thus, *univariate post hoc tests do not inform the researcher about the differences in the multivariate latent variables actually analyzed in the multivariate analysis.*

Understandably, Borgen and Seling (1978) argue:

> When data truly are multivariate, as implied by the application of MANOVA, a multivariate follow-up technique seems necessary to "discover" the complexity of the data. Discriminant analysis is multivariate; univariate ANOVA is not. (p. 696)

It is illogical to first declare interest in a multivariate omnibus system of variables, and to then explore detected effects in this multivariate world by conducting non-multivariate tests!

Examples of Bad Practice Within Dissertations. Three types of bad practice emerged as regards multivariate-related analyses in dissertations. First, *some students incorrectly employed univariate tests as post hoc tests following multivariate analyses.* For example, ##Horst (1991) reported that:

> The multivariate analysis of variance (MANOVA) indicated a significant [sic] difference between child molesters and the control group in family-of-

> origin characteristics ($F(2,59)=5.69$, $p<0.0055$, using Wilks' criterion). Separate univariate [post hoc] analyses indicated no significant [sic] effect for the measure of adaptability ($F(1,61)=1.58$, $p<0.2142$). (p. 30)

Secondly, *some students incorrectly interpreted their multivariate results*. For example, a very common error is using the multivariate lambda value to derive a test statistic, and to then incorrectly interpret the test as evaluating only a single multivariate effect or function.

In fact, lambda is an omnibus effect that is sensitive to all the effects or functions in a given analysis, and not to a single effect or function (Thompson, 1984, pp. 19-20). For example, in ##Morris's (1992) study, two perfectly uncorrelated multivariate effects or functions were computed across each of four groups of subjects. Yet, ##Morris (1992, pp. 101-102) incorrectly interpreted the test of lambda as only being a test of the first multivariate effect (i.e., the first canonical correlation coefficient) in each of four different sets of multivariate analyses.

##Morris (1992) also made a second, unrelated error. The student presumed that the first function or equation in each of the four sets of analyses were related to each other across the four groups. There was no basis whatsoever for this assumption.

Multivariate functions or equations are akin to factors in factor analysis. Across data sets the same functions, representing given constructs, functions, or factors, may appear in a different order within given analyses. One does not care that anxiety, for example, defines Function I in one group, but defines Function II in another group. What is generally important is that the same constructs appear across analyses, regardless of ordering.

##Morris (1992) merely presumed that Function I in all four analyses tapped the same construct. No evidence that these functions were comparable was presented. It is just as likely that across the four groups ##Morris (1992, pp. 101-102) was comparing the canonical correlation coefficients of incongruent functions or factors, i.e., comparing apples and prunes and tangerines rather than apples, apples, and apples.

But by far the most common error in the dissertations that were examined involved students *using univariate methods when multivariate methods were more appropriate*. The notions of Type I experimentwise error rate inflation, previously explored, bear upon this discussion.

##Stephenson (1992, pp. 47-49) reported 28 one-way ANOVAs, plus other statistical significance tests, all for a sample of $n=60$ cases of data. ##Thurstone (1991) was more ambitious, and conducted 60 t-tests. ##Burt (1991) was more ambitious still, and conducted 60 chi-square tests and 15 t-tests, all at $\alpha_{TW}=.05$.

But the lifetime prize for inflating Type I experimentwise error rate absolutely must go to ##Schmid (1991). ##Schmid's dissertation consisted of 277 pages, 143 (51.2%) of which alone

consisted of appended tables reporting univariate tests. ##Schmid's analyses included many tests done on data at the item level, i.e., with dependent variables consisting of responses to a single question (one wonders what the reliability of measures one item in length might have been). ##Schmid reported multi-way factorial ANOVAs and two multi-way factorial ANCOVAs for various single-item dependent variables and other dependent variables, all for a grand total of at least 1,017 univariate tests!

Principle #4: *Result interpretations should not be based only on standardized weights.*

All classical parametric analyses are correlational and use least squares weights, such as β weights, to optimize prediction (cf. Thompson, 1991a). However, many researchers do not recognize that such is the case, because most computer packages do not print the least squares weights that are actually invoked in ANOVA, for example, or when t-tests are conducted. Thus, some researchers unconsciously presume that such methods do not invoke optimal weighting systems.

Too many researchers presume that statistical packages only print results that are necessary for correct interpretation. Too many researchers presume that statistical packages print all the results that are necessary for correct interpretation.

Notwithstanding misconceptions to the contrary, all parametric analyses do invoke standardized weights similar to the beta (β) weights generated in regression. As Thompson (1992a) noted,

> These weights are all analogous, but are given different names in different analyses (e.g., beta weights in regression, pattern coefficients in factor analysis, discriminant function coefficients in discriminant analysis, and canonical function coefficients in canonical correlation analysis), mainly to obfuscate the commonalities of [all] parametric methods, and to confuse graduate students. (pp. 906-907)

If all standardized weights across analytic methods were called by the same name (e.g., beta weights), then students and other researchers might (correctly) conclude that all analyses are part of the same general linear model (Baggaley, 1981, p. 129; Bagozzi, 1981; Fornell, 1978, p. 168; Fan, 1992; Knapp, 1978).

A variable given a standardized weight of zero is being obliterated by the multiplicative weighting process, indicating either that (a) the variable has zero capacity to explain relationships among the variables or that (b) the variable has some explanatory capacity, but one or more other variables yield the same explanatory information and are arbitrarily (not wrongly, just arbitrarily) receiving all the credit for the variable's predictive power. On the other hand, as the standardized weights for variables deviate more from zero, these variables have more power to explain relationships among the variables.

Because a variable may be assigned a standardized multiplicative weight of zero when (b) the variable has some

explanatory capacity, but one or more other variables yield the same explanatory information and are arbitrarily (not wrongly, just arbitrarily) given all the credit for the variable's predictive power, it is *essential* to evaluate other coefficients in addition to standardized weights during interpretation, to determine the specific basis for the weighting. Just as it would be incorrect to evaluate predictor variables in a regression analysis only by consulting beta weights (Cooley & Lohnes, 1971, p. 55; Thompson & Borrello, 1985), it would be inappropriate in multivariate analyses to only consult standardized weights during result interpretation (Borgen & Seling, 1978, p. 692; Kerlinger & Pedhazur, 1973, p. 344; Levine, 1977, p. 20; Meredith, 1964, p. 55).

Examples of Bad Practice Within Dissertations. ##Stephenson (1992, p. 53) provides an example of bad practice in this area. ##Stephenson reported and interpreted only the standardized discriminant function coefficients from a discriminant function analysis. Absent the structure coefficients, it is impossible to know whether the variables in this study with smaller absolute standardized function coefficients (a) were relatively poorer predictors or (b) were arbitrarily deprived of credit for predictive power shared with other predictor variables.

But ##Gorsuch (1992) represents the most dramatic example of bad practice as regards misinterpreting only weights. In a multiple regression analysis, ##Gorsuch predicted a single dependent variable using scores on five predictor variables.

In Chapter IV ##Gorsuch (1992) incorrectly interpreted the standardized weights in the analysis, i.e., the β weights, as evaluating the relationship (i.e., the r) between the predictors and the dependent variable:

> Surprisingly, emotion script knowledge was negatively related [β = -.40] to sibling caregiving behavior. (p. 67)

Again, in Chapter V, ##Gorsuch (1992) repeated the same error:

> One of the most interesting findings of the study was that affective perspective-taking ability [β = +.26] but not emotion script knowledge [β = -.40] was positively associated with sibling caregiving behavior. (pp. 74-75)

The β weights in a regression analysis are the correlation coefficients between the respective predictors and the dependent variable only when those predictors that are correlated with the dependent variable are perfectly uncorrelated with each other. Such was not the case in this study.

Table 3 presents the regression coefficients in the manner in which they should have been presented by ##Gorsuch (1992, p. 66). The Table includes the structure coefficients so important in most data analytic situations (Thompson & Borrello, 1985). In regression analyses, to avoid result misinterpretation, both standardized weights and structure coefficients, or both standardized weights and correlation coefficients between the predictor variables and the dependent variable, should always be presented together.

INSERT TABLE 3 ABOUT HERE.

At a subsequent point, ##Gorsuch (1992) happily correctly inferred that this emotion script predictor variable had (a) the largest absolute value β weight and (b) the smallest absolute r with the dependent variable, because the variable was a pure suppressor:

> The negative association between emotion script knowledge and total sibling caregiving [β = -.40] should also be interpreted cautiously because it may be a statistical artifact of the regression analysis. Although emotion script knowledge was correlated with the predictor variables [sic], the zero-order correlation between emotion script knowledge and caregiving [r = .00] was not significant. This suggests that emotion script knowledge may have acted as a suppressor variable in the regression equation. (p. 76)

Less happily, ##Gorsuch (1992) seemed to think that this "statistical artifact" was peculiar, unusual, and perhaps therefore less noteworthy. In fact, suppression occurs with some frequency, and is an important and real dynamic within both reality and our analytic models (see Horst, 1966, p. 355; Pedhazur, 1982, p. 104).

Principle #5: *Intervally-scaled variables should generally not be converted to the nominal level of scale.*

In a seminal article, Cohen (1968, p. 426) noted that ANOVA and ANCOVA are special cases of multiple regression analysis, and argued that in this realization "lie possibilities for more relevant and therefore more powerful exploitation of research data." Since that time researchers have increasingly recognized that conventional multiple regression analysis of data as they were initially collected (no conversion of intervally scaled independent variables into dichotomies or trichotomies) does not discard information or distort reality, and that the "general linear model"

> ...can be used equally well in experimental or non-experimental research. It can handle continuous and categorical variables. It can handle two, three, four, or more independent variables... Finally, as we will abundantly show, multiple regression analysis can do anything the analysis of variance does--sums of squares, mean squares, F ratios--and more. (Kerlinger & Pedhazur, 1973, p. 3)

Discarding variance is not generally good research practice (Thompson, 1988b). As Kerlinger (1986, p. 558) explains,

> ...partitioning a continuous variable into a dichotomy or trichotomy throws information away... To reduce a set of values with a relatively wide range to a dichotomy is to reduce its variance and thus its possible correlation with other variables. A good rule of research data analysis, therefore,

> is: Do not reduce continuous variables to partitioned variables (dichotomies, trichotomies, etc.) unless compelled to do so by circumstances or the nature of the data (seriously skewed, bimodal, etc.).

Kerlinger (1986, p. 558) notes that variance is the "stuff" on which all analysis is based. Discarding variance by categorizing intervally-scaled variables amounts to the "squandering of information" (Cohen, 1968, p. 441). As Pedhazur (1982, pp. 452-453) notes,

> Categorization of attribute variables is all too frequently resorted to in the social sciences.... It is possible that some of the conflicting evidence in the research literature of a given area may be attributed to the practice of categorization of continuous variables.... Categorization leads to a loss of information, and consequently to a less sensitive analysis.

One reason why researchers may be prone (a) to categorizing continuous variables and also (b) to overuse of ANOVA is that some researchers unconsciously and erroneously associate ANOVA with the power of experimental designs. As Thompson (1993a) noted,

> Even most experimental studies invoke intervally scaled "aptitude" variables (e.g., IQ scores in a study with academic achievement as a dependent variable), to conduct the aptitude-treatment interaction (ATI) analyses recommended so persuasively by Cronbach (1957, 1975) in his 1957 APA Presidential address. (pp. 7-8)

Thus, many researchers employ interval predictor variables, even in experimental designs, but these same researchers too often convert their interval predictor variables to nominal scale merely to conduct OVA analyses.

It is *true* that experimental designs allow causal inferences and that ANOVA is appropriate for many experimental designs. However, it is *not* therefore *true* that doing an ANOVA makes the design experimental and thus allows causal inferences.

Humphreys (1978, p. 873, emphasis added) notes that:

> The basic fact is that a measure of individual differences is not an independent variable [in a experimental design], and it *does not become one* by categorizing the scores and treating the categories as if they defined a variable under experimental control in a factorially designed analysis of variance.

Similarly, Humphreys and Fleishman (1974, p. 468) note that categorizing variables in a nonexperimental design using an ANOVA analysis "not infrequently produces in both the investigator and his audience the illusion that he has experimental control over the independent variable. Nothing could be more wrong." Since all analyses are correlational, and it is the design and not the analysis that yields the capacity to make causal inferences, the

practice of converting intervally-scaled predictor variables to nominal scale so that ANOVA and other OVAs (i.e., ANCOVA, MANOVA, MANCOVA) can be conducted is inexcusable, at least in most cases.

As Cliff (1987, p. 130, emphasis added) notes, the practice of discarding variance on intervally scaled predictor variables to perform OVA analyses creates problems in almost all cases:

> Such divisions are not infallible; think of the persons near the borders. Some who should be highs are actually classified as lows, and vice versa. In addition, the "barely highs" are classified the same as the "very highs," even though they are different. Therefore, reducing a reliable variable to a dichotomy [or a trichotomy] makes the variable <u>more unreliable</u>, not less.

In such cases, it is the reliability of the dichotomy that we actually analyze, and <u>not</u> the reliability of the highly-reliable, intervally-scaled data that we originally collected, which impact the analysis we are actually conducting.

<u>Examples of Bad Practice Within Dissertations</u>. The seemingly automated mutilation of intervally-scaled independent or predictor variables was fairly common within the two cohorts of Ph.D. dissertations examined here. Three patterns emerged within these studies.

First, *some students employed sample-specific centiles as cutpoints to create equal-sized groups, and were apparently oblivious to the fact that such practices limit result generalizability across studies*. When researchers working within an area all use their own sample-specific medians, for example, to create balanced group sizes, then usually each and every researcher is using a different cutoff score to define group membership.

Then the results across these studies are no longer directly comparable. Where researchers obtain divergent results, the divergence may be an artifact of using different cutoffs. Where findings are similar across studies, again one does not know if the similarities across results are artifactual or real. This makes the cumulation of knowledge across such studies very difficult.

> ##Fisher (1992) provides an example of these bad practices: For this analysis, low machismo scores were defined as those at or below the 50th percentile; high scores were those above the 50th percentile. (p. 113)

##Kaiser (1991) provides another example, noting that "Median splits were performed on three of the hypotheses" (p. 98).

Second, *where cutoffs other than centiles are employed, the cutoffs should be explicitly justified on some empirical or theoretical basis*. ##Cattell (1992) provides an example of the violation of this principle. Scores on the 30-item, dichotomously-scored, Geriatric Depression Scale/Amended were used in the study, and potentially ranged from 0 to 30. ##Cattell (1992) decided, without any explicit justification, that "A score of 11 or above placed subjects in the depressed group" (p. 45).

Perhaps other researchers have consistently used this score as

a cutoff. Perhaps the score was the median in a normative sample of the general population. Perhaps there was a theoretical rationale for the choice. We simply don't know why this choice was made, and therefore what the implications of the choice may be.

Third, and most importantly, *the impacts of converting scales of measurement on the score reliability (and other features) of the data are underrecognized within dissertation research*. Some of these impacts have been previously described.

##Mulaik (1992) provides an example. It was most commendable that ##Mulaik (1992) computed some reliability estimate for the scores in that sample:

> A reliability analysis was computed for the total scores from the Christian Religiosity Scale [in the present study]. The standardized item alpha was computed to be .945. (p. 58)

But then ##Mulaik converted these generally reliable scores into a dichotomy for the purposes of substantive analyses.

> This [present] study also used a [sample-specific] median split to classify the students into high or low Christian religiosity categories. (p. 51)

The reliability of the newly mutilated scores was no longer .945, and the reliability of the scores actually analyzed, i.e., the mutilated scores, was unreported by ##Mulaik (1992).

Systematic variance is what makes scores reliable. Scores with more total variance tend to have more systematic variance. Mutilating intervally-scaled scores into nominal categories usually reduces total variance, usually by a lot, and thus also usually mutilates the reliability of the scores that are actually analyzed.

But taking a variable with very little total variance and apparently little reliability, and then trichotomizing the variable, makes the least sense of all. ##Lawley (1992) took scores on a variable named PNID, with a mean of 1.18 and a standard deviation of 0.70 (p. 59), and reported that:

> PNID scores of .65 or below were included in the bottom third while scores of 1.09 or above were included in the top third. (p. 55)

It is especially interesting that the highest score on this variable in ##Lawley's (1992) study was apparently 3.43 (p. 57). As ##Lawley (1992) acknowledged, the PNID authors themselves recommend a cutoff score of 4 for classifying subjects as being severely depressed. Thus, the <u>highest</u> score in ##Lawley's (1992) entire sample appeared to be <u>less than the minimum</u> cutoff score suggested by the test's own authors!

<u>An Ancillary Comment on the Value of Confession</u>. Dissertation authors apparently sometimes feel that confession of methodological errors either absolves guilt or at least is emotionally cathartic. For example, ##Kaiser (1991) noted that:

> These [median splits for three hypotheses] are of course less refined than other techniques.... Now that significant [sic] relationships between social anxiety and other constructs have been uncovered for children [using mutilated variables], use of

> procedures such as regression analyses [in future studies without mutilated variables] should increase what is known about these relationships. (p. 98)

##Mulaik (1992) expressed similar sentiments:

> There were no differences found between those who rated high on the Christian Religiosity Scale and those who rated low. This result was rather surprising.... The failure to find significance [sic] with this variable may be due to the fact that in this study it was [measured at interval level but then mutilated and] used as a dichotomous variable and significance [sic] may have been found if the variable was [kept] continuous. (p. 74)

While acknowledging what would be better practice is nice, acknowledging one's errors while at the same time committing them seems somewhat disingenuous. Confession during the act of committing the error is not a reasonable substitute for avoiding the error itself.

Principle #6: *Covariance corrections are generally either unnecessary or ineffective, and should therefore usually be avoided.*

Although ANCOVA is used by some researchers even in studies lacking randomized assignments to groups, empirical studies of research practice indicate that ANCOVA is not frequently employed. This is partly because important ANCOVA methodological assumptions are most likely to be met when researchers do random experimental assignment, and random experimental assignment is rare (Welch & Walberg, 1974, p. 113). Analysis of covariance (ANCOVA) has been used in about four percent of the recently published research (Elmore & Woehlke, 1988; Goodwin & Goodwin, 1985; Willson, 1980).

ANCOVA does accurately adjust for pretreatment group differences, but only conditionally--when important methodological assumptions are met. Huitema (1980) and Loftin and Madison (1991) present accessible summaries of the relevant conditions.

What ANCOVA Actually Is. As explained elsewhere (Thompson, 1992b), conceptually, ANCOVA first "residualizes" the dependent variable of all the variance that is linearly predictable with the covariate variable(s). Then the resulting "error" or "e" scores are used as the new dependent variable in an ANOVA.

The way the residualization is accomplished is by first using regression to predict the dependent variable with the covariate(s), *completely ignoring*, for the moment, the fact that subjects may be in different groups or cells. Thus, Figure 1 portrays the relationship between a dependent variable and a single covariate, and the figure does not invoke the concept of groups.

INSERT FIGURE 1 ABOUT HERE

Regression analysis employs two types of weights: an additive constant ("a") applied to every case and a multiplicative constant

("b") applied to the predictor variable score (X_i) for each of the i subjects. This yields a predicted variable score for each person, $\hat{Y}_i$, that is the optimal prediction of each subject's actual score on the dependent variable, Y_i. Thus, the weighting system takes the form of a regression equation:

$$Y_i \text{ <---- } \hat{Y}_i = a + b\ (X_i)$$

The error of the prediction for each person, e_i, is:

$$e_i = Y_i - \hat{Y}_i$$

As can be seen by examining Figure 1, since their areas do not overlap, for a given data set the e and the X scores are always perfectly uncorrelated ($r = 0$), and so are the e and the $\hat{Y}$ scores. Conceptually, *ANCOVA is nothing more that an ANOVA done on these e scores*, i.e., on the residualized $\hat{Y}$ scores.

The "Homogeneity of Regression" Assumption. It would be wonderful if this "statistical correction" for pre-existing group differences could always be used. Some researchers incorrectly believe that ANCOVA has just such magic, and

> can "save" a shoddy experiment [with major, real, pre-existing group differences]. Some researchers overuse this method as in the instance of a person I once overheard asking of a researcher, "Where is your analysis of covariance?"--the understanding in his department was that it is always used in experimentation. (McGuigan, 1983, p. 231)

Unfortunately, there is no more magic in statistics than there is other aspects of life. If the groups are different (e.g., a compensatory education group and a group not eligible for compensatory education) at the start of a study, ANCOVA cannot always be used to statistically adjust for these pre-existing differences.

As might be logically expected, what is required to use a single regression equation to compute the e scores is that this single equation is a *reasonable* one for the subjects in each of the groups or cells in the study, considered separately. More specifically, what is actually important is that the B weights for the covariate, computed separately in each group, are reasonably comparable. Statisticians call this the "homogeneity of regression" assumption, because this phrase sounds fancier than saying simply the "equality of the B weights" assumption. Quite simply, for a single covariate situation it is only reasonable to use a single B weight to compute e scores for the subjects in all the groups if the B weights in each group are about the same and, thus, about the same as the single B weight computed ignoring group membership.

The consequences of failing to meet this assumption will be discussed momentarily. However, it is worth noting that, unless the groups were created through random assignment, the B weight relationships between the covariate and the dependent variable often are not equivalent across the groups.

For example, when the covariate is an achievement pretest score, and the dependent variable is an achievement posttest, the

regression equation drawn in the scattergram for a group of subjects actually represents the learning curve for the group. If one group consists of children eligible for a compensatory intervention based on low pretest scores, and the comparison group consists of children not eligible for the intervention, we would not normally expect their pretest averages or their learning curves to be comparable. Thus, when we most wish to have statistical magic to equate divergent groups, that is exactly when the ANCOVA correction is least likely to be useful.

ANCOVA uses a single equation that may differ from the regression lines of all the groups in the analysis when the assumption of homogeneity of regression is not met. For example, if the regression line slopes upward at a 75 degree angle for one group, and upward at a 25 degree angle for the other group, and an average or 50 degree regression line (i.e., equation) is used for both groups, both groups' dependent variables scores will be "corrected" inappropriately, because a 50 degree regression line is incorrect for both groups. Too few researchers understand the consequences of such inappropriate ANCOVA corrections.

In the 1963 Handbook of Research on Teaching, Campbell and Stanley wrote an influential chapter on experimental and quasi-experimental design. Campbell and Stanley (1963, p. 193) suggested that "the use of this more precise analysis [e.g., ANCOVA] would seem highly desirable." They also argued that "covariance analysis and blocking on 'subject variables' such as prior grades, test scores, parental occupation, etc., can be used, thus increasing the power of the significance test" (p. 196).

Campbell and a colleague subsequently issued what appeared to be a recant noting that the decision to blithely use statistical control when the homogeneity of regression assumption is not met leads to "tragically misleading analyses" that actually "can mistakenly make compensatory education look harmful" (Campbell & Erlebacher, 1975, p. 597). Similarly, Cliff (1987, p. 273) argues that, "It could be that the relationship between the dependent variable and the covariate is different under different treatments. Such occurrences tend to invalidate the interpretation of the simple partial correlations described above."

The "Statistical Power" Issue. Some researchers argue that ANCOVA increases statistical power against Type II error, by reducing the "error" portion of the depending variable, without changing the variance (sums-of-squares) attributable to the other independent variables (i.e., the ways or factors in the study). This happens, but only when the covariate is correlated with the dependent variable and is uncorrelated (hopefully perfectly uncorrelated) with the independent variables.

When the covariate is related to the treatment variable, use of the covariance correction will alter the effects attributed to the treatment itself. For example, one might have a very effective intervention that looks completely ineffectual, because the covariate is given credit for the variance that would correctly otherwise be attributed to the treatment variable. Here the ANCOVA correction actually destroys power against Type II error.

Using Multiple Covariates. Some researchers believe that using multiple covariates is OK, or even desirable. This is classical "more is always better" thinking. Unfortunately, there are problems with this "thinking".

Actually, when there are multiple covariates, the regression equation for the statistical adjustment simply has more predictors, and associated B weights. But there is an inherent dilemma in using covariance corrections, especially when multiple covariates are used. The problem is conceptual, and is far too infrequently recognized because sometimes researchers don't think reflectively about their analytic choices, and miss the forest for seeing the trees.

Put simply, covariance corrections may result in the analysis of a dependent variable that is no longer interpretable. As Thompson (1992b, pp. xiii-xiv) notes, "Statistical corrections remove parts of the dependent variable, and then analyze whatever's left [i.e., the e scores], even if whatever's left no longer makes any sense. At some point we may no longer know what it is we're analyzing". As Thompson (1991b, p. 508) suggests,

> Consider an actual [reading] dissertation (see Thompson, 1988[a]) in which the posttest achievement variable was "corrected" using four pretest achievement subtests. What was the posttest achievement variable after this correction?... [W]hatever it was, this student probably wasn't analyzing achievement after this nuclear weapon covariance correction.

ANCOVA analyzes e scores, not observed Y scores. It becomes increasingly hard to interpret the e scores we're analyzing as we employ more covariates. As Cliff (1987, p. 278) explains, "since this [statistical correction] is really a form of regression, inferences become slipperier as the variables [covariates] increase" in number. Here, more is usually not better, and even one may be too much.

Examples of Bad Practice Within Dissertations. ANCOVA was used in several of the Ph.D. dissertations examined here. For example, ##Huberty (1991) reported that,

> State anxiety was covaried in order to remove it's [sic] influence. (p. 67)

Similarly, ##Kaiser (1991) reported that,

> The social desirability by LOC [locus of control] ANCOVA resulted in a significant [sic] main effect on social anxiety for social desirability. (p. 78)

##Spearman (1991) went for the prize, and performed four MANCOVAs and three ANCOVAs, one with two covariates.

##Lawley (1992), like most students, employed ANCOVA with groups that were not created using random assignment. ##Lawley (1992) explained,

> To control for differences in academic ability, multivariate analysis of covariance (MANCOVA) was performed, with achievement variance removed. The pattern of results was unchanged. (p. 61)

##Schmid (1991) rationalized this analytic selection thusly:

> The purpose of the covariates was to increase the statistical power in analyzing the findings.... (p. 49)

The importance of the ANCOVA assumptions has been previously explained. Neither ##Huberty (1991), not ##Kaiser (1991), nor ##Spearman (1991), nor ##Lawley (1992), nor ##Schmid (1991), explicitly evaluated whether they meet this assumption.

Principle #7: *Stepwise methods should not be used.*

Stepwise methods are used with some frequency in behavioral research, but perhaps most frequently in Ph.D. dissertations. Stepwise discriminant analysis, and especially stepwise multiple regression analysis, both appear to be popular.

But there are serious problems with these methods, at least in almost all applications. These are summarized in Snyder (1991), in Huberty (1989), and in Thompson's (1989b) editorial titled, "Why won't stepwise methods die?". Therefore, these problems are only briefly summarized here. This summary is couched in the context of a regression analysis, but the discussion fits equally well to other stepwise applications.

Stepwise methods typically are implemented in a so-called forward selection mechanism. A researcher has a set of predictors. The predictor that explains the most variance in the criterion variable (i.e., has the largest r^2 with the criterion variable scores) is entered in the first step of the analysis.

In the next step of analysis, the predictor next entered is not necessarily the predictor with the second largest r^2 with the criterion variable. Rather, the general question then addressed at each step is, "which additional single predictor will explain the largest portions of the criterion variable's variance, *not counting any variance already explained by previously entered predictors*?".

For example, predictor X_1 might have the largest r^2 with Y, i.e., $r2 = 50\%$. Predictor X_2 might have the second largest r^2 with Y, i.e., $r2 = 49\%$. Predictor X_3 might have the smallest r^2 with Y, i.e., $r2 = 5\%$. Predictor X_2 would definitely be entered in the first step of the analysis. But if the 49% of the variance explained by X_2 is all within the 50% of the Y variance already explained by X_1, X_2 will not be entered in the second step. If X_3 explains any the Y variance that is unexplained by X_1, then X_3 will be entered in the second step of analysis.

There are three major problems with conventional applications of these methods. First, *stepwise methods do not correctly identify the best set of predictors of a given predictor variable set size, k*. For example, if one has 30 predictors variables, and does three steps of analysis, it is possible that the best predictor set of size $k=3$ will include none of the three variables selected after three steps of stepwise analysis of the same data.

This may seem counter-intuitive, but upon reflection, it should be easy to see that in fact stepwise analysis does not seek to identify the best predictor set of a certain size. Stepwise

simply does not ask the question, "What is the best predictor set of a given size?" This question requires simultaneously considering all the combinations of the variables that are possible for a given set size. Stepwise analysis never *simultaneously* considers all the combinations of the predictor variables. Rather, at each step stepwise analysis takes the previously entered variables as a given, and then asks which *one* change in the predictor set will most improve the prediction.

Second, *stepwise methods tend to yield results that are sample-specific and do not generalize well to future studies*. This is because stepwise requires a linear sequence of decisions, each of which is contingent upon the previous decisions in the sequence. This is very much like walking through a maze--an incorrect decision at any point will then lead to a cascade of subsequent decisions that each may themselves be wrong.

Stepwise considers all differences of any magnitudes between variance explained by the predictor variables to be exact and true. Since there are usually numerous combinations of the predictor variables, and credit for variance explained for each partition of the predictors may be influenced by sampling error, any small amount of sampling error anywhere in a single predictor variable can lead to disastrous choices in the linear sequence of stepwise selection decisions.

In the previous example, perhaps even though all other results in the sample are exactly true in the population, as reflected in the true population parameters, perhaps X_2 in the population explains 50.000001% of the variance in Y. The variable that in the sample won't be entered at all should have been entered first. This may be why Cliff (1987, pp. 120-121) suggested that, "a large proportion of the published research results using this method probably present conclusions that are not supported by the data."

Third, *most computer packages (and therefore most doctoral students) employ the incorrect degrees of freedom in their statistical significance tests for stepwise methods, thus systematically always inflating the likelihood of obtaining statistically significant results*. Degrees of freedom are like coins that we can spend to investigate what's going on within our data, i.e., what explains or predicts the variability in the dependent variable. The total number of coins we have to spend within regression is called the degrees-of-freedom total, and equals n-1.

Regression partitions these coins into two parts: the portion we've spent to get answers to questions about what explains the dependent variable, and the portion that we haven't yet spent. The first of these two partitions is called the degrees-of-freedom (df) explained (or, to confuse graduate students, by any of the synonymous terms, df regression, df model, or df between). The second of these two partitions is called the degrees-of-freedom (df) unexplained (or, to confuse graduate students, by any of the synonymous terms, df residual, df error, or df within). The df explained equals the number of predictor variables (p) "used". The

df error equals n-1 - p.

The question becomes, what does "used" mean. The computer packages define "used" as actually entered into the prediction equation. Thus, if n was 100, and there were 50 predictors variables, but only three steps of analysis done, df total would be 99, df explained according to the computer packages would be 3, and df error would be 96.

However, in this example each and every one of the 50 predictor variables was "used" at each and every one of the three steps, to decide which predictor to enter at each step. The 47 predictors may have been returned to cafeteria, but each one was examined, and played with, eaten, and used, prior to the return to the display case. The df explained at the third step (and at every step) for this example should have been 50, while the df error should have been 49.

It is instructive to see how using the wrong degrees of freedom in the numerator of the statistical significance testing calculations, and the wrong denominator df in the calculations, both bias the tests in favor of getting statistical significance. These dynamics are illustrated for this example within Table 4.

INSERT TABLE 4 ABOUT HERE.

Clearly, using the wrong degrees of freedom in both the numerator and the denominator can outrageously affect statistical significance tests. No wonder Cliff (1987, p. 185) says that, "most computer programs for [stepwise] multiple regression are positively satanic in their temptation toward Type I errors"!

Examples of Bad Practice Within Dissertations. Several Ph.D. dissertations within these two cohorts employed stepwise methods. Stepwise method's tendencies toward Type I errors are worse as the number of variables is larger while the number of steps is smaller, and/or as the number of subjects is smaller. Thus, ##Stephenson's study is intriguing; ##Stephenson (1992, p. 51) performed a stepwise discriminant analysis involving 21 predictor variables and 48 cases.

Table 5 presents the incorrect degrees of freedom, the F calculated, and the p calculated reported by ##Velicer (1992, p. 50) for a stepwise analysis. The p calculated value was <.00001, so p calculated was less than α_{TW}, and the null hypothesis was rejected. Table 5 also presents the correct statistics, not reported in the dissertation, for this study. The p calculated value actually was .09471, i.e., the results were not really statistically significant at conventional alpha levels.

INSERT TABLE 5 ABOUT HERE.

But ##Fisher's (1992, p. 128) results are also dramatic. Both the reported incorrect results and the unreported correct results are presented in Table 6. The reported p calculated value was .00602, resulting in statistical significance, while the correct p

calculated value was .19005, not resulting in statistical significance.

INSERT TABLE 6 ABOUT HERE.

Discussion

It must be said that even a methodologically flawed dissertation might still contribute to the literature. But the problem with methodologically flawed studies is that these methodological flaws are gratuitous. There is no excuse for bad methodological practice in dissertations.

Maybe a student cannot afford to hire the Gallup organization to draw a national probability sample, or maybe informed consent will not be given by enough people to allow an ideally-designed study to be done, or maybe an ideal piece of equipment to acquire certain measurements cannot be afforded. But these practical considerations do not bear upon analytic choices--the student is fully in control. Of course, such gratuitous errors are especially peculiar when students confess their errors, while committing them, as against avoiding these errors.

Dissertations are a critical component of the knowledge creation endeavor. Under the cruel-and-usual punishment clause of the U.S. Constitution, people may only be subjected to writing a dissertation once in their lives. Only this once is the expertise of a doctoral candidate linked with the combined expertise of four or five or six dissertation committee members, each with their own terminal degree. Dissertations are also a relatively unique form of scholarship in that page limits do not constrain depth of exploration or breadth of coverage. Thus, scholars care about dissertation quality, and it is unfortunate when academic integrity is gratuitously compromised.

References

Baggaley, A. R. (1981). Multivariate analysis: An introduction for consumers of behavioral research. Evaluation Review, 5, 123-131.

Bagozzi, R.P. (1981). Canonical correlation analysis as a special case of a structural relations model. Multivariate Behavioral Research, 16, 437-454.

Borgen, F.H., & Seling, M.J. (1978). Uses of discriminant analysis following MANOVA: Multivariate statistics for multivariate purposes. Journal of Applied Psychology, 63(6), 689-697.

Campbell, D. T., & Erlebacher, A. (1975). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In M. Guttentag & E. L. Struening (Eds.), Handbook of evaluation research (Vol. 1, pp. 597-617). Beverly Hills: SAGE.

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage, Handbook of research on teaching (pp. 171-246). Chicago: Rand McNally.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt Brace Jovanovich.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Cooley, W.W., & Lohnes, P.R. (1971). Multivariate data analysis. New York: John Wiley & Sons.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L.J. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.

Dawis, R.V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

Denton, J.J., Ciou-Yeuh, T., & Chevrette, P. (1988). Quality of research experience in graduate programs as perceived by faculty, graduates and current students. National Forum of Applied Educational Research Journal, 1(1), 23-29.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Eason, S., & Daniel, L. G. (1989, January). Trends and methodological practices in several cohort dissertations. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 306 299)

Edgington, E.S. (1964). A tabulation of inferential statistics used in psychology journals. American Psychologist, 19, 202-203.

Edgington, E.S. (1974). A new tabulation of statistical procedures in APA journals. American Psychologist, 29(1), 25-26.

Elmore, P.B., & Woehlke, P.L. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17(9), 19-20.

Fan, Xitao. (1992, April). Canonical correlation analysis as a general data-analytic model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 348 383)

Fish, L.J. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Fornell, C. (1978). Three approaches to canonical analysis. Journal of the Market Research Society, 20, 166-181.

Goodwin, L.D., & Goodwin, W.L. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14(2), 5-11.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Horst, P. (1966). Psychological measurement and prediction. Belmont, CA: Wadsworth.

Huberty, C. (1989). Problems with stepwise methods--better alternatives. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 43-70). Greenwich, CT: JAI Press.

Huberty, C.J, & Morris, J.D. (1989). Multivariate analysis versus multiple univariate analysis. Psychological Bulletin, 105, 302-308.

Huck, S.W, & Cormier, W.G. (in press). Reading statistics and research (2nd ed.). New York: Harper Collins.

Huitema, B.E. (1980). The analysis of covariance and alternatives. New York: John Wiley & Sons.

Humphreys, L.G. (1978). Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis. Journal of Educational Psychology, 70, 873-876.

Humphreys, L.G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. Journal of Educational Psychology, 66, 464-472.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart and Winston.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart and Winston.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kupfersmid, J. (1988). Improving what is published: A model in

search of an editor. American Psychologist, 43, 635-642.
Lagaccia, S.S. (1991). Methodology choices in a cohort of education dissertations. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 149-158). Greenwich, CT: JAI Press.
Levine, M. S. (1977). Canonical analysis and factor comparison. Newbury Park, CA: SAGE.
Loftin, L., & Madison, S. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 133-147). Greenwich, CT: JAI Press.
Love, G. (1988, November). Understanding experimentwise error probability. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville. (ERIC Document Reproduction Service No. ED 304 451)
McGuigan, F. J. (1983). Experimental psychology: Methods of research (4th ed.). Englewood Cliffs: Prentice-Hall.
Maxwell, S. (1992). Recent developments in MANOVA applications. In B. Thompson (Ed.), Advances in social science methodology (Vol. 2, pp. 137-168). Greenwich, CT: JAI Press.
Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.
Meier, S.T., & Davis, S.R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. Journal of Counseling Psychology, 37, 113-115.
Meredith, W. (1964). Canonical correlations with fallible data. Psychometrika, 29, 55-65.
Moore, M.A. (1991, April). The place of significance testing in contemporary social science. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 333 036)
Morrow, J.R., Jr., & Frankiewicz, R.G. (1979). Strategies for the analysis of repeated and multiple measures designs. Research Quarterly, 50, 297-304.
Nunnally, J.C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.
Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart and Winston.
Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.
Reinhardt, B. (1991, January). Factors affecting coefficient alpha: A mini Monte Carlo study. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio. (ERIC Document Reproduction Service No. ED 327 574)
Reinhardt, B. (1992, April). Estimating result replicability using double cross-validation and bootstrap methods. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.(ERIC Document Reproduction Service No. ED 346 164)
Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its

display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.
Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.
Rowley, G.L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.
Schmid, J. & Leiman, J. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 53-61.
Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation Review, 8, 573-582.
Shaver, J. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61(4), 293-316.
Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), (1991). Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 99-105). Greenwich, CT: JAI Press.
Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61(4), 334-349.
Tatsuoka, M. M. (1973). Multivariate analysis in educational research. In F. N. Kerlinger (Ed.), Review of research in education (pp. 273-319). Itasca, IL: Peacock.
Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Newbury Park: SAGE.
Thompson, B. (1986, November). Two reasons why multivariate methods are usually vital. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis.
Thompson, B. (1987, January). Peer review of doctoral dissertations as a quality control mechanism: Some methods and examples. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas. (ERIC Document Reproduction Service No. ED 282 499)
Thompson, B. (1988a, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)
Thompson, B. (1988b). Discarding variance: A cardinal sin in research. Measurement and Evaluation in Counseling and Development, 21, 3-4.
Thompson, B. (1989a). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.
Thompson, B. (1989b). Why won't stepwise methods die?. Measurement and Evaluation in Counseling and Development, 21(4), 146-148.
Thompson, B. (1991a). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling

and Development, 24(2), 80-95.
Thompson, B. (1991b). [Review of Data analysis for research designs]. Educational and Psychological Measurement, 51, 500-510.
Thompson, B. (1992a). DISCSTRA: A computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function and structure coefficients and group centroids. Educational and Psychological Measurement, 52, 905-911.
Thompson, B. (1992b). Misuse of ANCOVA and related "statistical control" procedures. Reading Psychology, 13(1), iii-xviii.
Thompson, B. (1992c). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.
Thompson, B. (1993a, April). The General Linear Model (as opposed to the classical ordinary sums of squares) approach to analysis of variance should be taught in introductory statistical methods classes. Paper presented at the annual meeting of the American Educational Research Association, Atlanta. (ERIC Document Reproduction Service No. ED 358 134)
Thompson, B. (1993b). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377.
Thompson, B. (1994, January). It is incorrect to say "The test is reliable": Bad language habits can contribute to incorrect or meaningless research conclusions. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.
Thompson, B. (in press). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality.
Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.
Welch, W. W., & Walberg, H. J. (1974). A course evaluation. In H. J. Walberg (Ed.), Evaluating educational performance: A sourcebook of methods, instruments, and examples (pp. 113-124). Berkeley: McCutchan.
Wick, J.W., & Dirkes, C. (1973). Characteristics of current doctoral dissertations in education. Educational Researcher, 2, 20-22.
Willson, V.L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.
Witte, R.S. (1985). Statistics (2nd ed.). New York: Holt, Rinehart and Winston.

Table 1
Formula for Estimating Experimentwise Type I Error Inflation
When Hypotheses are Perfectly Uncorrelated

TW alpha		Tests		Experimentwise alpha
1 - (1 - 0.05)	**	1	=	
1 - (0.95)	**	1	=	a
1 - 0.95			=	0.05000
Range Over Testwise (TW) alpha = .01				
1 - (1 - 0.01)	**	5	=	0.04901
1 - (1 - 0.01)	**	10	=	0.09562
1 - (1 - 0.01)	**	20	=	0.18209
Range Over Testwise (TW) alpha = .05				
1 - (1 - 0.05)	**	5	=	0.22622
1 - (1 - 0.05)	**	10	=	0.40126
1 - (1 - 0.05)	**	20	=	0.64151
Range Over Testwise (TW) alpha = .10				
1 - (1 - 0.10)	**	5	=	0.40951
1 - (1 - 0.10)	**	10	=	0.65132
1 - (1 - 0.10)	**	20	=	0.87842

Note. "**" = "raise to the power of".

'These calculations are presented (a) to illustrate the implementation of the formula step by step and (b) to demonstrate that when only one test is conducted, the experimentwise error rate equals the testwise error rate, as should be expected if the formula behaves properly.

Table 2
All Possible Families of Outcomes
for a Fair Coin Flipped Three Times

	Flip # 1	2	3	
1.	T	T	T	
2.	H	T	T	p of 1 **or more** H's (TW error analog)
3.	T	H	T	in set of 3 Flips = 7/8 = 87.5%
4.	T	T	H	
5.	H	H	T	or
6.	H	T	H	where TW error analog = .50,
7.	T	H	H	EW p = 1 - $(1 - .5)^3$
8.	H	H	H	= 1 - $(.5)^3$
				= 1 - .125 = .875
p of H on each Flip	50%	50%	50%	

Note. The probability of one or more occurrences of a given outcome in a set of events is 1 - $(1-p)^k$, where p is the probability of the given occurrence on each trial and k is the number of trials in a set of perfectly independent events.

Table 3
Regression Coefficients for the ##Gorsuch (1992) Example

Predictor	β [a]	r_s [b]	r_s^2 [b]
cognition	0.15	0.405	0.164
emotion	-0.40	0.000	0.000
affective	0.26	0.490	0.240
care	0.32	0.618	0.382
age	0.26	0.341	0.116

[a]Coefficients reported by ##Gorsuch (1992, p. 66).
[b]Coefficients not reported by ##Gorsuch (1992).

Table 4
Example Stepwise Results
Illustrating the Bias Toward Type I Error

Incorrect Computer Version

Source	Sum of Squares	df	Mean Squares	Fcalc	pcalc
Explained	20	3	6.6667	8.0000	0.000125
Unexplained	80	96	0.8333		
Total	100	99			

Correct Version

Source	Sum of Squares	df	Mean Squares	Fcalc	pcalc
Explained	20	50	0.4000	0.2300	0.999999
Unexplained	80	46	1.7391		
Total	100	99			

Table 5
Correct and Incorrect Test Statistics for the ##Velicer Study

Incorrect Values Reported by ##Velicer

SOS	df	MS	Fcalc	pcalc
40	1	40	38.66666	<.00000001
60	58	1.034482		
100	59			

Correct Values Not Reported by ##Velicer

SOS	df	MS	Fcalc	pcalc
40	17	2.352941	1.647058	=.09471154
60	42	1.428571		
100	59			

Table 6
Correct and Incorrect Test Statistics for the ##Fisher Study

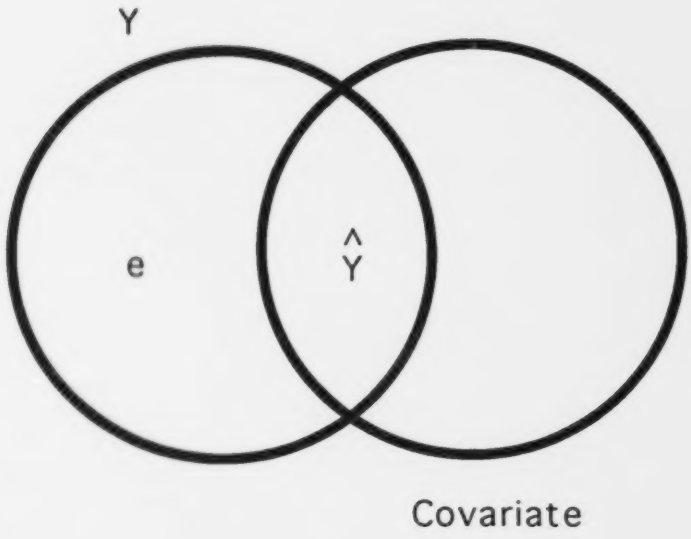
Incorrect Values Reported by ##Fisher

SOS	df	MS	Fcalc	pcalc
5.5	1	5.5	7.798941	0.0060231686
94.5	134	0.705223		
100	135			

Correct Values Not Reported by ##Fisher

SOS	df	MS	Fcalc	pcalc
5.5	5	1.1	1.513227	0.1900507810
94.5	130	0.726923		
100	135			

Figure 1
How Dependent Variable (Y) and Covariate Relationships
Create Scores on Latent Synthetic Variables e and $\hat{Y}$



Doctoral programs that wish to survive both accreditation and coordinating board reviews can ill afford to ignore the quality of their students' dissertations. Review teams invariably examine dissertations as a key, if not the key, reflection of the cumulative impacts of programs' doctoral training and mentorship.

Thus, various strategies for improving the methodological quality in dissertations have been proposed (cf. Thompson, 1987). And reviews of good and bad practices in a program's dissertations may themselves be helpful to the program (Denton, Ciou-Yeuh & Chevrette, 1988; Eason & Daniel, 1989).

The purpose of the present paper is to review common methodology errors made within dissertations. Thus, the present paper does something different than characterizing typical practice. Reviews of typical analytic practice are already available as regards both published research (Edgington, 1964, 1974; Elmore & Woehlke, 1988; Goodwin & Goodwin, 1985; Willson, 1980) and dissertation research (cf. Lagaccia, 1991; Wick & Dirkes, 1973). The present paper employs the same format of a previous study within the same genre (Thompson, 1988a), and the previous study could be examined to extrapolate some general trends.

To make the discussion of common methodology mistakes in dissertations more concrete, specific actual examples of methodology errors are cited. These examples were derived from Ph.D. studies completed within one department at a large Research I university during the calendar years, 1991 and 1992. The department houses two APA-accredited psychology programs, in addition to other programs.

However, to minimize embarrassment to these former students (or perhaps to their dissertation committee members), pseudonyms are employed as citations to these dissertations. The pseudonyms are differentiated by the use of pound signs as part of these citations (e.g., ##Lawley, 1992).

This discussion is organized within the framework of seven analytic principles. Each principle is explained, and then illustrative examples of violations of the principles from Ph.D. dissertations are presented.

No effort has been made to cite all the errors within the dissertations studied. Nor does the present paper cite errors that fall outside the framework of the seven analytic principles.

For example, students within these two cohorts may have experienced problems in following the APA style guide, especially as regards (a) the use of ampersands in citations in narrative, as against within parentheses, (b) the correct use of "et al.", and (c) the correct ordering of multiple citations within a single parenthetical list. ##Lawley (1992) provides one example involving the first two considerations:

> As a criterion measure, the CDI has been used in a number of childhood depression research studies (Hughes, [sic] et al., 1990; Kaslow, et al., 1984; Lobovits & Handal, 1985; McCauley, et al., 1988; Mullins, et al., 1985,; Scwartz, et al., 1982; Worchel, Hughes, et al., 1990).... Carey, Faulstich,